Australian Bureau of Statistics

1351.0.55.056 - Research Paper: A Statistical Framework for Analysing Big Data, Jun 2015

Latest ISSUE Released at 11:30 AM (CANBERRA TIME) 30/06/2015 First Issue

Summary

Executive Summary

EXECUTIVE SUMMARY

In this paper, it is contended that the threshold challenges that must be adequately addressed before Big Data sources can be used for the production of official statistics are the business case, the validity of statistical inference, and data ownership and access issues.

The business case comprises business needs and benefits, and data ownership and access issues are particularly important where, as is commonly the case, the National Statistical Office is not the custodian of the Big Data source. Above all, given the expected inferential biases from Big Data – due to under-coverage, self-selection, missing values etc. – statistical methods must be developed before Big Data sources can be harnessed for the production of official statistics.

Using a Bayesian framework, this paper outlines necessary conditions – in particular, the Missing At Random condition – for valid statistical inference to be made for estimating or predicting finite population parameters (e.g. totals of population units), or for estimating the super-population parameters of statistical models (e.g. the regression coefficients of a linear regression model).

By assuming that Missing At Random conditions are fulfilled, the paper also provides an illustrative theoretical method for utilising satellite imagery data to predict crop areas and crop yields. The analysis assumes that the data are described by a dynamic logistic model for crop types and a dynamic linear model for crop yields. The method relies on using “ground truth” data from a random sample to calibrate the satellite imagery, and using the latter as covariates to predict the data of interest for the population not included in the random sample.

Finally, the paper outlines methods to address related statistical computing issues and proposes strategies for extending the model to provide a better fit to the observed data.

About this Release

In this paper, it is contended that the threshold challenges that must be adequately addressed before Big Data sources can be used for the production of official statistics are the business case, the validity of statistical inference, and data ownership and access

issues.
Using statistical modelling, the paper outlines necessary conditions for addressing the biases inherent in Big Data sources when estimating parameters of a finite population or super-population model.
To illustrate the proposed statistical framework, the paper describes a method, based on State Space modelling, for utilising satellite imagery data to predict crop types and crop yields. The paper also outlines methods to address related statistical computing issues, and proposes strategies for extending the model to provide a better fit to the observed data.

© Commonwealth of Australia

All data and other material produced by the Australian Bureau of Statistics (ABS) constitutes Commonwealth copyright administered by the ABS. The ABS reserves the right to set out the terms and conditions for the use of such material. Unless otherwise noted, all material on this website – except the ABS logo, the Commonwealth Coat of Arms, and any material protected by a trade mark – is licensed under a Creative Commons Attribution 2.5 Australia licence